The Relationship Between Traffic Volume and Auto Accidents

Jinhui Zhang

A Master's Project Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science Department of Mathematics

ILLINOIS STATE UNIVERSITY   SUMMER 2017

Abstract

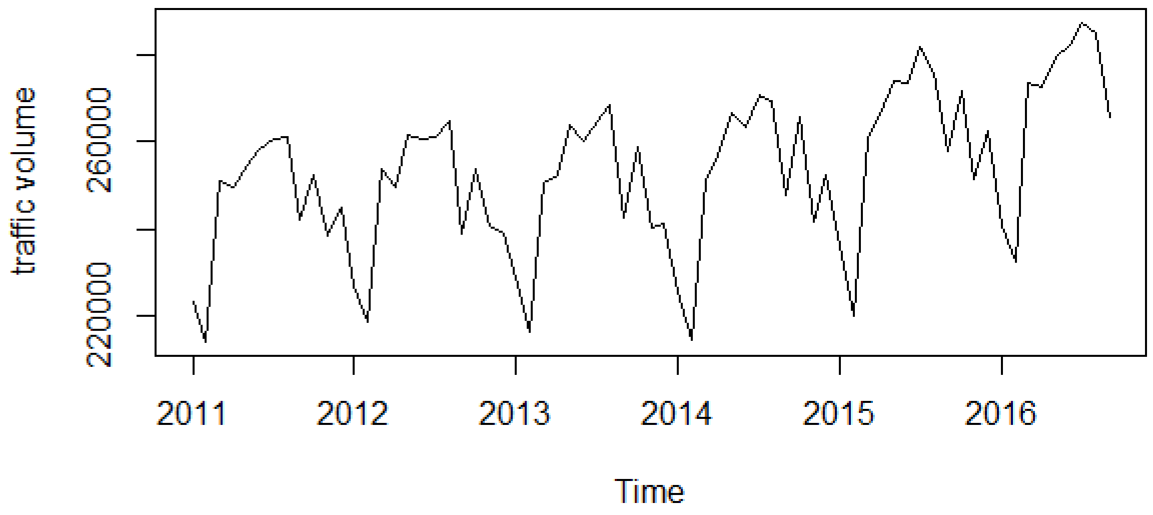
As number of cars running on the road increasing recent decade, the car accidents has been more or less effected from frequency, severity and other aspects. This study is trying to find if there is any relationship between the traffic volume and car accidents, taking distracted driving into consideration as well assuming increasing traffic volume may increase the chance of distraction. Incurred losses is used as a representative measurement of severity and frequency of the auto accidents base on the definition of loss cost. Regression time series data is used to analysis and reach a conclusion.

Chapter 1: Background of the study

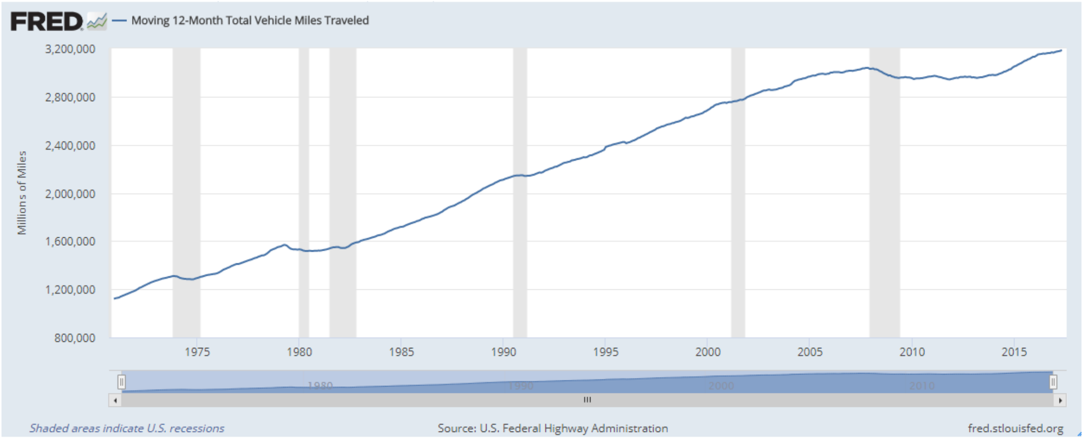
Introduction

For insurance agencies, everyone is very happy seeing the traffic volume is growing, which mean they are getting a growing opportunities to sell more policies. But according to underwriting and actuarial work, the risk of losses for each individual vehicle can not be simply adding up. The concern is, if more cars on the roads have effect on the overall loss potentially will occur, or it does not have an effect to be worry about.

We can see a slowly increasing and seasonal trend in traffic volume according to the data from U.S Department of Transportation Federal Highway Administration, Office of Highway Policy Information:

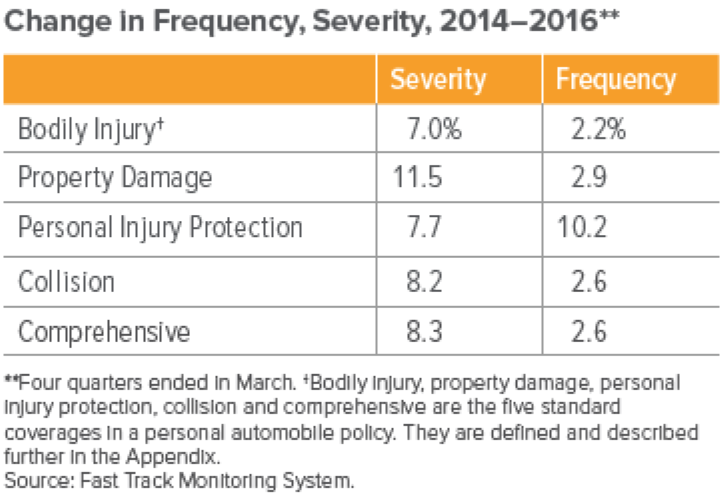


And an increasing in yearly total vehicle miles traveled



Statement of the problem

Base on *Personal Automobile Insurance – More Accidents, Larger Claims Drive Costs Higher* (Insurance Information Institute, October 2016), auto accidents are increasing both in size and frequency. Collision increased 8.2% in severity and 2.6% in frequency.



Change in Frequency, Severity, 2014–2016**

	Severity	Frequency
Bodily Injury†	7.0%	2.2%
Property Damage	11.5	2.9
Personal Injury Protection	7.7	10.2
Collision	8.2	2.6
Comprehensive	8.3	2.6

**Four quarters ended in March. †Bodily injury, property damage, personal injury protection, collision and comprehensive are the five standard coverages in a personal automobile policy. They are defined and described further in the Appendix.
Source: Fast Track Monitoring System.

The Loss Cost of auto insurance also increased in different types of claims.

The article given the following potential cause of the increasing costs:

1. People are driving more than before in mileage and the traffic volume.
2. Distracted driving – more and more people use electronic devices while driving.
3. Increasing size of average claim (severity)

Then the article gives a few suggestions to drivers what they can do to lower their auto insurance rates, but we are not further study here.

Rationale for the study

As more and more auto vehicles running on the road, which potentially also means more cars get insured, one of the major contain according to insurer would be if this increasing traffic volume has anything to do with the property and injury damages cause by auto accidents, as well as any potential factor that is related to the losses caused by car accidents.

This study is to find out if the increasing loss trend is related to the three reasons giving as above.

Using the Loss Cost to measure both frequency and severity is based on the definition of Loss Cost:

$$\text{Loss Cost} = \text{frequency} \times \text{severity} \div 100$$

Here frequency is the number of claims per 100 vehicle-years. Severity is the average size of a claim (dollar amount).

Research questions or hypothesis

This research trying to see if there is a relationship between traffic volume and the car accidents.

There are two measures of auto accidents, frequency and severity, two variables can work together be represented by loss cost.

Hypothesis: The relationship could be a bell curve shape since at the beginning as the number of vehicles running on the road increase, it would be easier to have accidents, this is when the amount of car doesn't have effect on the speed driving. At this point, more cars create more distraction so the lost cost goes up. But as the traffic volume keep increasing, it will eventually have negative effect on the overall speed people driving, at this point the accidents are less likely to happen when cars drive too low, can't go anywhere (imagine sitting in a traffic jam). Both accident rate and severity would go down and the loss cost going down.

So we expect a curve (of loss cost) going up first then going down as the traffic volume going up.

Assumptions, limitations and delamination

Since we focus on the car accidents, so we are only interested in the losses caused by auto incidents, we should take bodily injury, property damage and collision into consideration as they are relevant to accidents, while the comprehensive category of losses with not be considered.

This study looking at the two variable on a quarterly scale throughout years. Due to lack of data resource, we are not able to manipulate the situation only looking at one specific road to get traffic volume and car accidents information on a "micro scale".

Definitions of terms

Definition of severity: Severity is the average size of a loss.

Definition of frequency: Frequency is the average number of claims per time period, usually per year.

Incurred loss is in four categories,

Body injury: covers another person's medical expenses if the insured person cause an accident.

Property damage: covers the damage the insured person cause to another person's property in a car accident.

Body injury and property damage belongs to liability coverage.

Collision Coverage helps pay to repair or replace the insured person's vehicle if it's stolen or damaged from any kinds of collision.

Comprehensive pays to repair or replace the insured person's vehicle if it's damaged by things like hall, animal damage or everything else not collision.

Chapter 2: Review of literature

Prior research

A study in 2003 by Lena Hiselius (Lund University, Sweden) studied the relationship between the number of vehicles per hour and the accident frequency. The study discussed the distribution of the number of accidents in a certain time interval.

Poisson distribution is commonly used in this kind of study, based on the total number of accidents is reasonably stable over a certain time interval while each individual accidents are independent.

The study decided to use Negative Binomial distribution as an extension of Poisson distribution for the number of accidents occurs on a specific road.

The advantage of this study is it collected data on road sections, the number of passing vehicles was from Swedish National Road Administration, and collision data was from police report. And the data contained different road types.

The study applied both Poisson and Negative Binomial regression model, and found out they both a good fit.

The number of passing traffic is a counted number, and this is when it is good to use Poisson regression when the response variable Y is a count. And we can also have $\frac{Y}{t}$, the rate or incidence as the response variable, where t is an interval representing time. So this is why in the study in 2003, Poisson regression can be used.

In our case, since the response variable would be the amount of money caused by car accidents, Poisson regression won't apply.

Related literature

Stationarity.

A time series is stationary when it is "stable", meaning:

1. The mean is constant over time (no trend)
2. The correlation structure remains constant over time.

There are two methods being used to test is a time series' stationarity, Box-Ljung method and ADF method.

Stationarity has very good properties, given data $x_1, x_2, \ldots, x_n$ we can estimate by averaging.

For example, if the mean is constant, we can estimate it by the sample average $\bar{x}$.

Pairs can be used to estimate correlation on different lags:

Such as $(x_1, x_2), (x_2, x_3), (x_3, x_4), \ldots$ for lag 1.

$(x_1, x_3), (x_2, x_4), (x_3, x_5), \ldots$ for lag 2.

When a time series is trend stationary, it will have stationary behavior around a trend.

This model can be expressed by $Y_t = \alpha + \beta t + X_t$ where X_t is stationary.

A different type of model for trend is random walk, which has the form $X_t = X_{t-1} + W_t$ where W_t is white noise. It is called a random walk because at time t the process is where it was at time $t-1$ plus a completely random movement. For a random walk with drift, a constant is added to the model and will cause the random walk to drift in the positive or negative direction of the drift.

If data is not stationary, time series often generated as $X_t = (1 + p_t)X_{t-1}$, meaning that the value of the time series observed at time t equals the value observed at time $t-1$ and a small percent change p_t at time t. Typically, p_t is referred to as the return or growth rate of a time series, and this process is often stable. It can be shown that the growth rate p_t can be approximated by

$Y_t = logX_t - logX_{t-1} \approx p_t$.

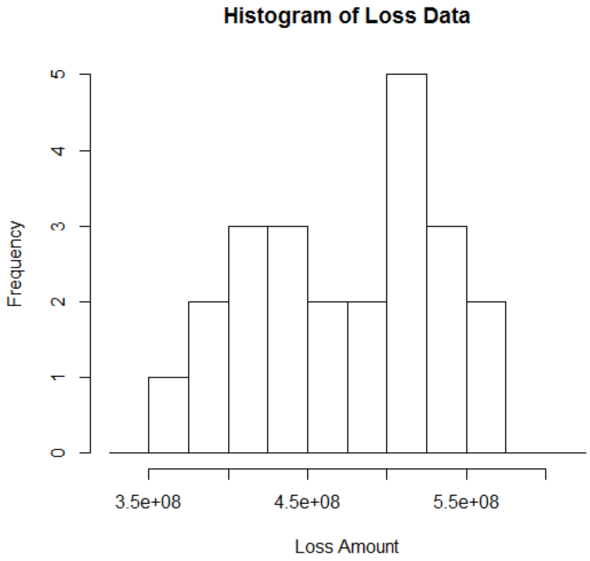
Chapter 3: Methodology

Development of instruments

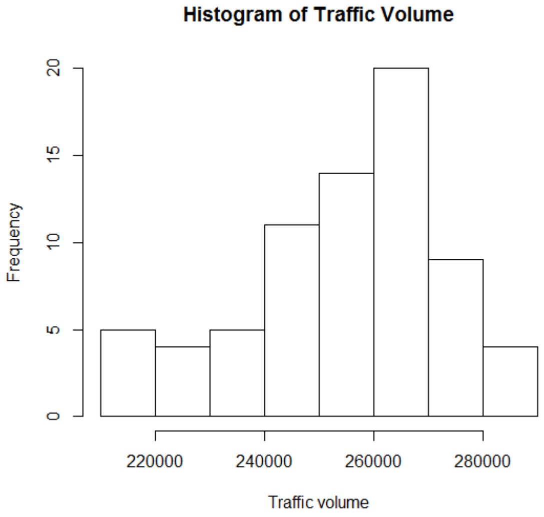
To study the relationship between the traffic volume and the accidents, we are trying to look at in the insurance prospect of view, if the losses will keep increasing, and in what speed the losses are increasing. Furthermore, since distraction has become one of the major cause of car accidents, we will also take this into consideration as well.

First take a look at the distribution of the traffic volume and the losses. They are expected to be approximately normal distributed.

Histogram graph



Histogram graph of traffic volume

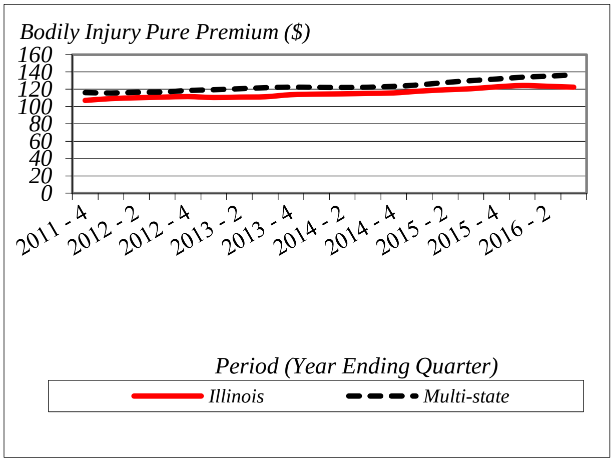


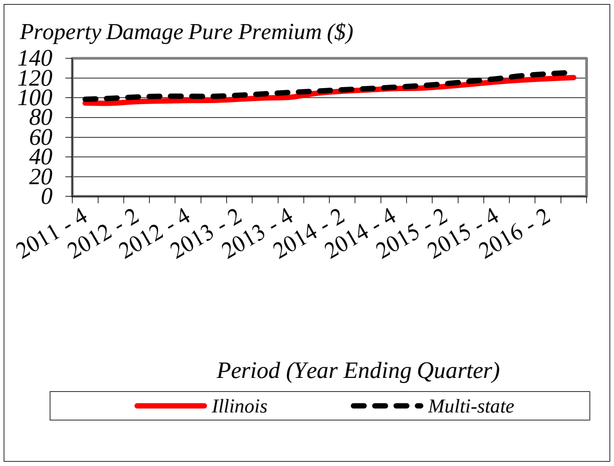
Data description

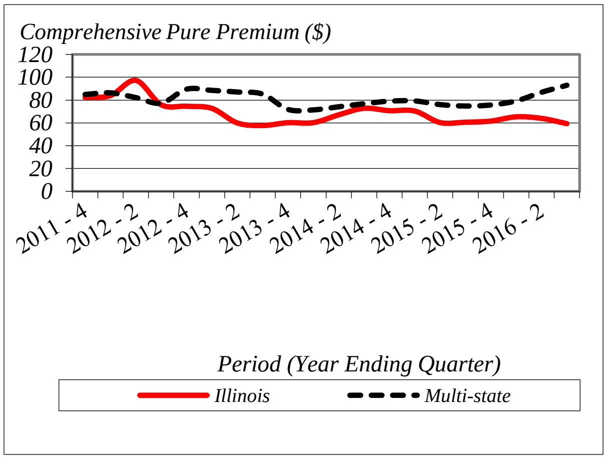
The traffic volume data is monthly data from Federal Highway Administration. We will also convert monthly data into quarterly data to be consistent with the loss data.

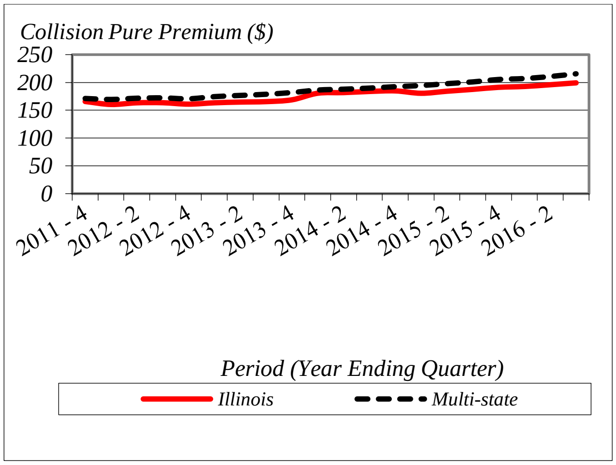
The loss data comes from Fast Track Plus, which contains 2011 to 2016 quarterly loss data in bodily injury, property damage, comprehensive and collision. We are using the incurred loss of bodily injury, property damage and collision as representative of Loss Cost variable, which both frequency and severity are taken into consideration.

Fast Track Plus is a product created by Independent Statistical Service, Inc., a subsidiary of the Property Casualty Insurers Association of American and enhanced by Pinnacle Actuarial Resources, Inc. It provides the information that insurers need to make competitive underwriting decisions – earned premium, pure premium, loss ratio, etc. It also provided graphs that display trends and compare individual state experience to countrywide figures (as below).









We can tell from the graph and the definition of "Comprehensive" data is not closely related to the auto accidents and it will add randomness in finding out the relationship, so this is why this category data would not be used in this study.

The data used in this study is only about Illinois state, consider Illinois state is following the trend of the big picture. Multi-state includes all of the states in the United States and regional data inside and outside United States.

Distraction data is also from Federal Highway Administration, provided data in percentage annually from year 2009 to 2015. The percentage shows in what percentage people were distracted among the driver involved in accidents happened in the past years national. It also provided the original numbers of crashes, include overall crashes and the number of crashes affected by distraction.

Distraction data is annually data. In order to match all different data to be quarterly, we average yearly data into quarterly and assume the crashes uniformly distributed in each quarter.

Research design

Regression time series data.

Traffic volume and loss data are both time series data has an increasing trend. In order to apply regression with time series data, remove the trend of the data by differencing the data until the time series data is stationary.

Then perform the regression with same orders time series data.

The Ljung-Box test and Augmented Dickey-Fuller (ADF) methods are used to check stationary.

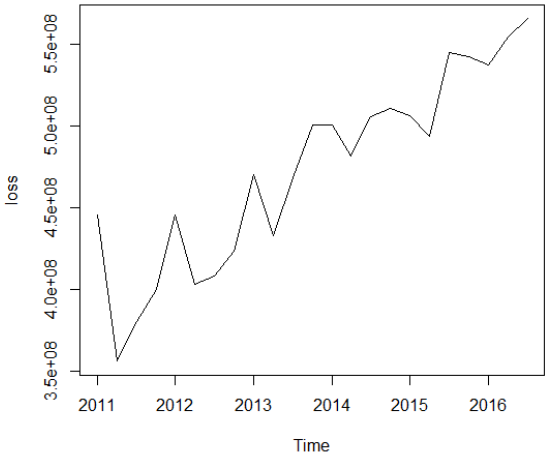
The Ljung-Box examines whether there is significant evidence for non-zero correlations at lags 1-20. Small p-value (less than 0.05) suggest that the series is stationary.

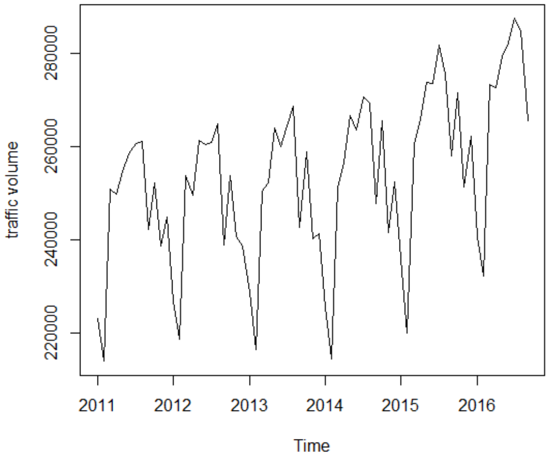
In the Augmented Dickey-Fuller t-statistic test, small p-values suggest the data is stationary and doesn’t need to be differenced.

Due to the small adjusted R-square value, we add in another factor – distracted driving after simply considering traffic volume and auto accidents losses by themselves.

Chapter 4: Analysis of Data

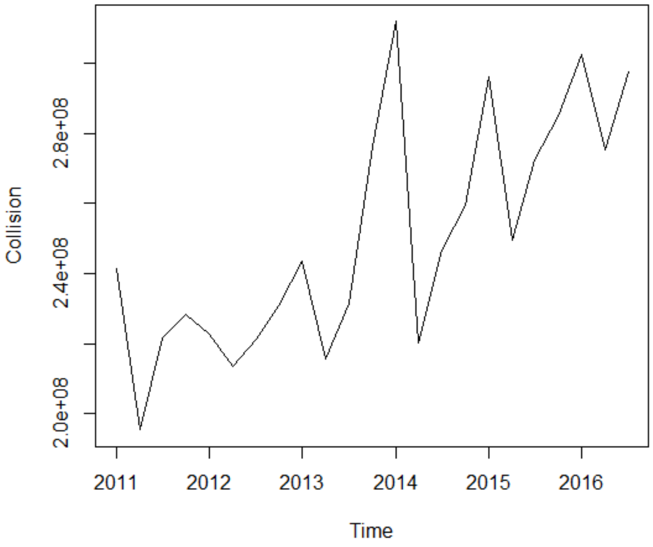
Observe increasing trend



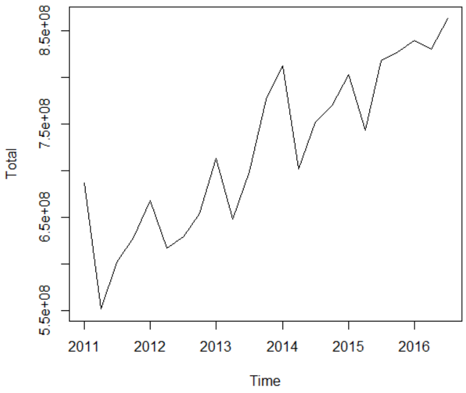


Here the loss data does not include the “collision” category. This “loss” data contain “body injury” and “property damage” – insurance liability.

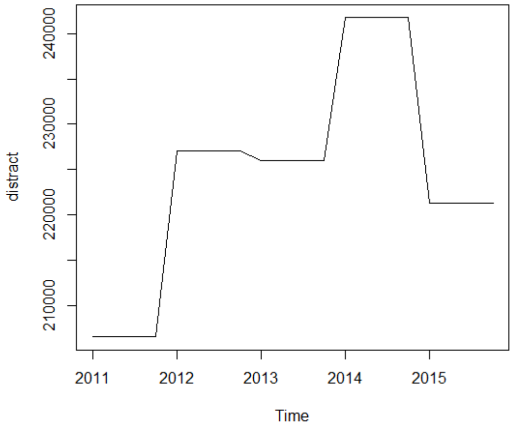
Looking at “collision” category separately,



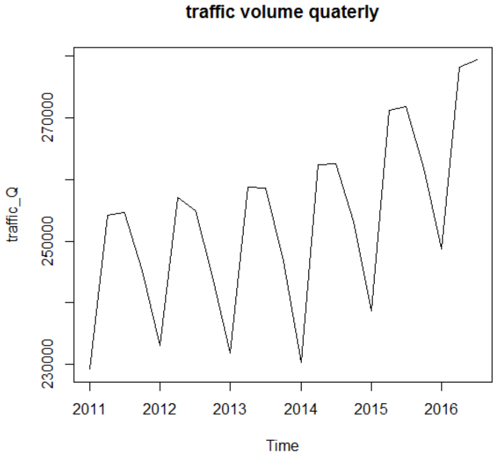
Combine collision and liability loss together we have graph as below,



And last we have the distracted data average into quarterly,



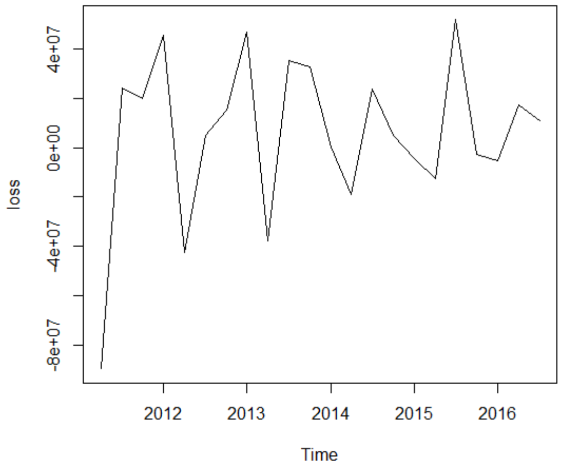
Also convert traffic volume monthly data into quarterly using the average of three months in a season.

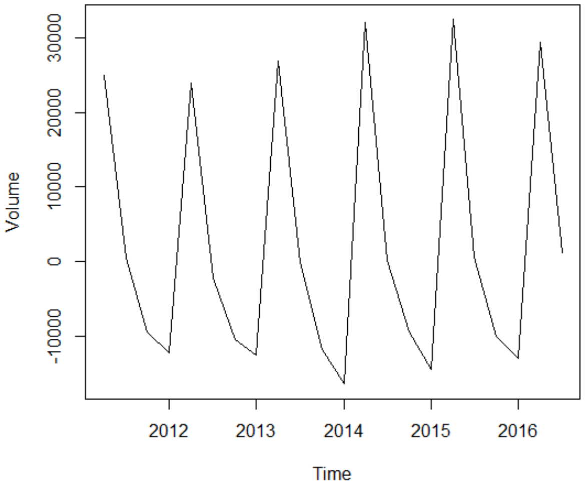


Now all of the data are in quarterly form, the next step is to remove the increasing trend to obtain stationary data.

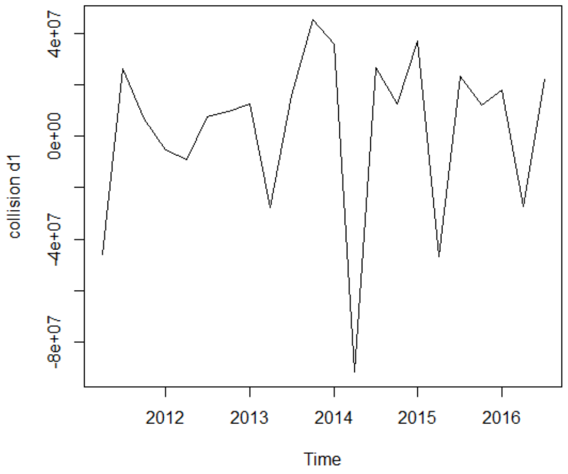
Remove the trend by differencing,

Differencing once,

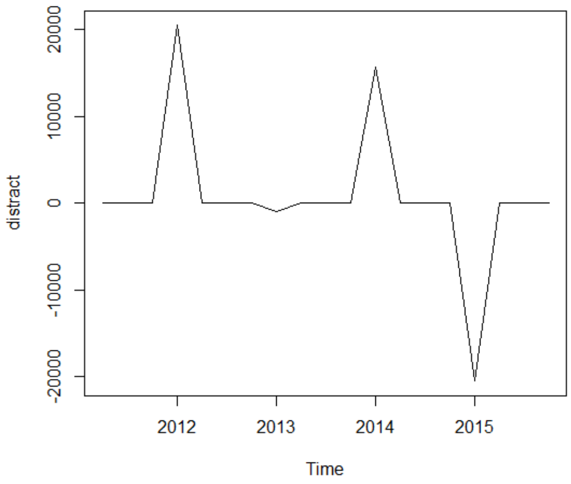




first difference collision



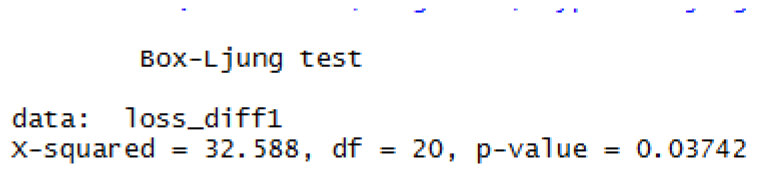
first difference distracted crash



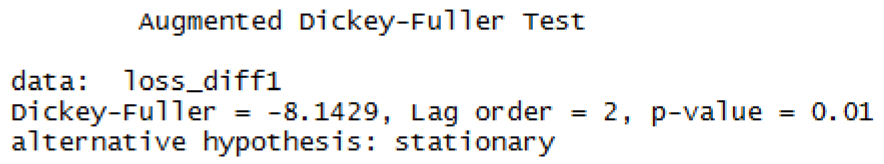
After the first differencing, we want to test if they are stationary, if they are, then they are ready to use. If they are still not stationary, we will keep differencing until they are stationary.

Test stationary.

For loss first difference.

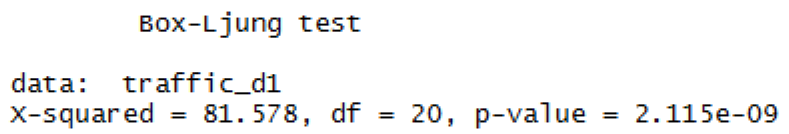


```
        Box-Ljung test

data:  loss_diff1
X-squared = 32.588, df = 20, p-value = 0.03742
```

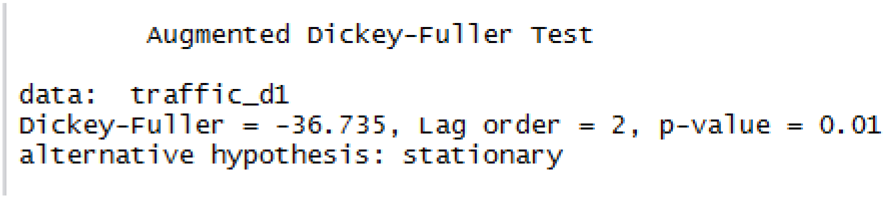


```
        Augmented Dickey-Fuller Test

data:  loss_diff1
Dickey-Fuller = -8.1429, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

For traffic volume first difference.

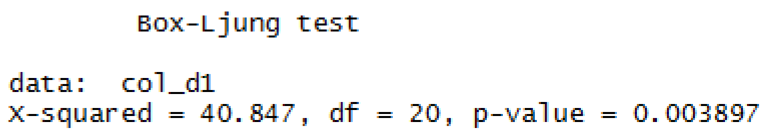


```
        Box-Ljung test

data:  traffic_d1
X-squared = 81.578, df = 20, p-value = 2.115e-09
```

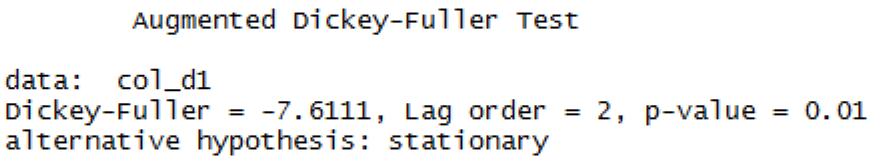


```
        Augmented Dickey-Fuller Test

data:  traffic_d1
Dickey-Fuller = -36.735, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

For collision first difference.



```
        Box-Ljung test

data:  col_d1
X-squared = 40.847, df = 20, p-value = 0.003897
```



```
        Augmented Dickey-Fuller Test

data:  col_d1
Dickey-Fuller = -7.6111, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

Since the p-values less than 0.05, they are stationary based according to both test methods, no more differencing needed.

Distracted driving crashes first difference is not stationary base on the two tests, we will consider about this later.

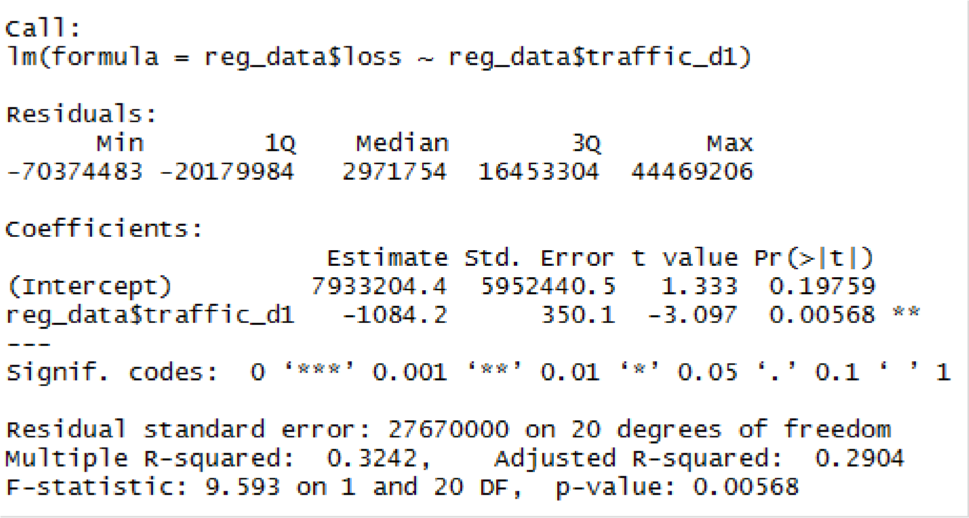
Regression.

First regression model is between liability loss (body injury and property damage) and traffic volume.

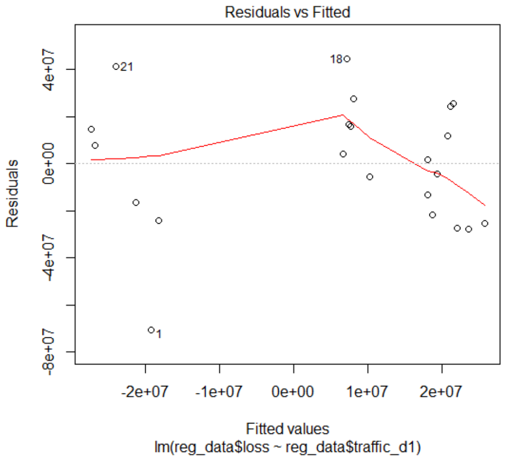
From the outcome of the regression analysis, we can see the p-value 0.00568 is reasonably small.

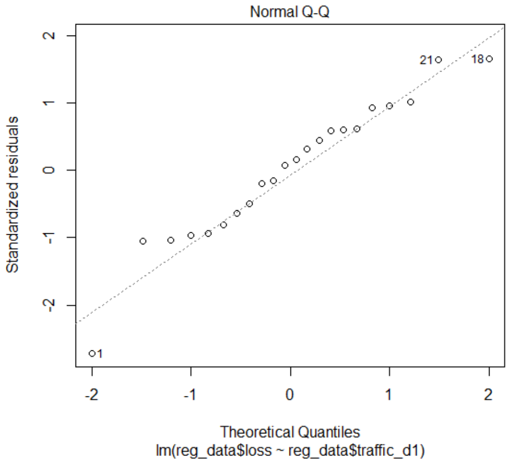
The variables has a correlation with 2 stars significance, which is good.

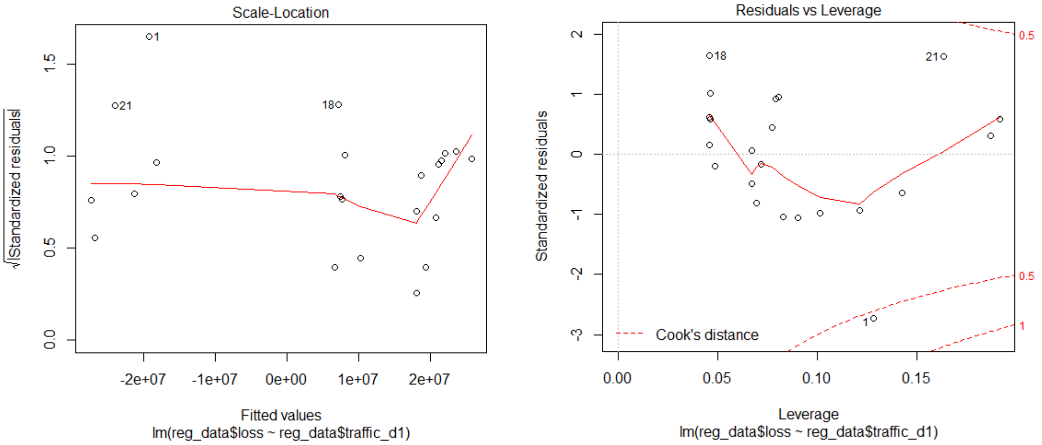
From the residual plot, the residuals are approximately normally distributed but not so convincing, they are evenly spread above and below 0, except there are three outliers at point 1, 18 and 21.



```
Call:
lm(formula = reg_data$loss ~ reg_data$traffic_d1)

Residuals:
      Min        1Q    Median        3Q       Max 
-70374483 -20179984   2971754  16453304  44469206 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)   
(Intercept)         7933204.4  5952440.5   1.333  0.19759   
reg_data$traffic_d1   -1084.2      350.1  -3.097  0.00568 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27670000 on 20 degrees of freedom
Multiple R-squared:  0.3242,	Adjusted R-squared:  0.2904 
F-statistic: 9.593 on 1 and 20 DF,  p-value: 0.00568
```







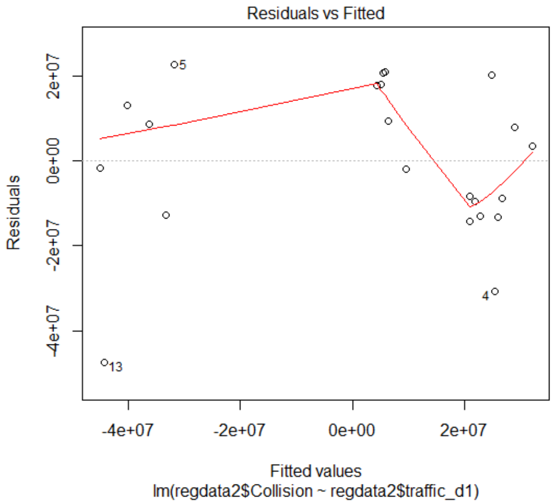
The Adjusted R-squared provided by regression summary is 0.29, which is quite small, along with small p-value, indicates the model fit well but will not be precise if we use this model for prediction. It means that additional explanatory variables may be needed. So, we should take distracted driving into consideration.

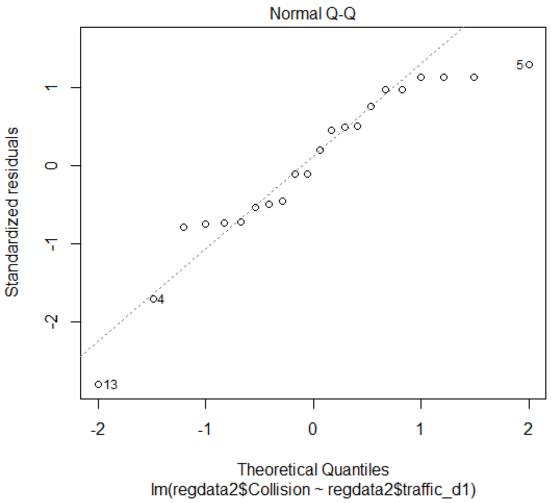
After removed the three outliers and fit regression model again, the new regression has an adjusted R-squared 0.4 and similar other parameters as the original regression. And the new regression suggested another three outliers. So removing the "outliers" from the original regression is not worth doing, the first three outliers should be kept in the model.

Second regression model is between Collision loss amount and traffic volume.

```
Call:
lm(formula = regdata2$Collision ~ regdata2$traffic_d1)

Residuals:
      Min        1Q    Median        3Q       Max 
-47388186 -12078281    841103  16622207  22578187 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6127726.6  4021659.6   1.524    0.143    
regdata2$traffic_d1   -1574.5      236.5  -6.657 1.76e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18690000 on 20 degrees of freedom
Multiple R-squared:  0.6891,	Adjusted R-squared:  0.6735 
F-statistic: 44.32 on 1 and 20 DF,  p-value: 1.758e-06
```

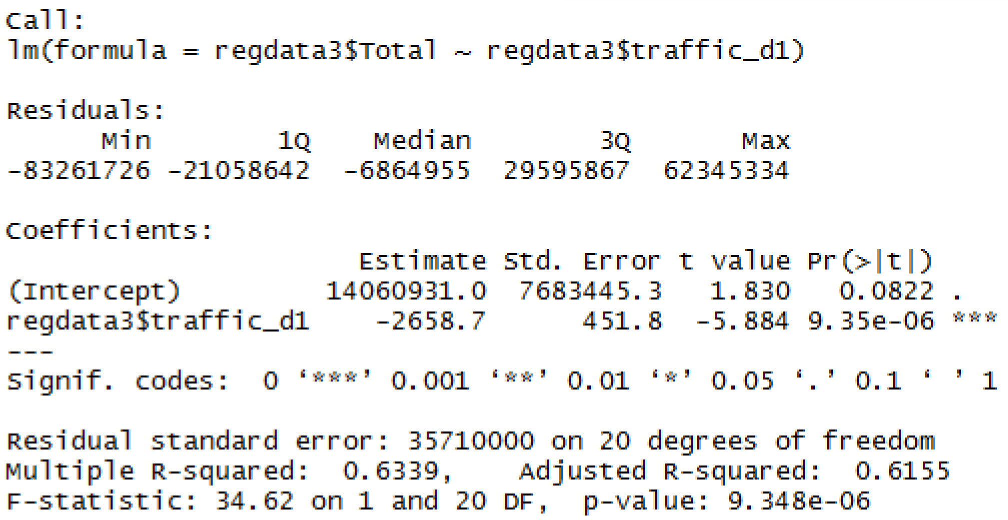
Using collision loss data, the traffic volume has 3 stars significance and the adjusted R-squared become 0.67 compare with 0.29 when using liability loss amount.





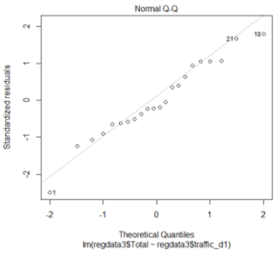
But from the Q-Q plot, residuals still not perfectly normally distributed.

Once we combine all of the loss amount together (include both liability and collision loss), the result is not much different from the result of only collision. For most part collision loss acting as a major influence in the regression model.



```
Call:
lm(formula = regdata3$Total ~ regdata3$traffic_d1)

Residuals:
      Min        1Q    Median        3Q       Max 
-83261726 -21058642  -6864955  29595867  62345334 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         14060931.0  7683445.3   1.830   0.0822 .  
regdata3$traffic_d1    -2658.7      451.8  -5.884 9.35e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35710000 on 20 degrees of freedom
Multiple R-squared:  0.6339,	Adjusted R-squared:  0.6155 
F-statistic: 34.62 on 1 and 20 DF,  p-value: 9.348e-06
```

We can see adding liability losses has a slightly negative effect on adjusted R-squared – from 0.67 to 0.61 and also adding a little bit significance to the intercept parameter.

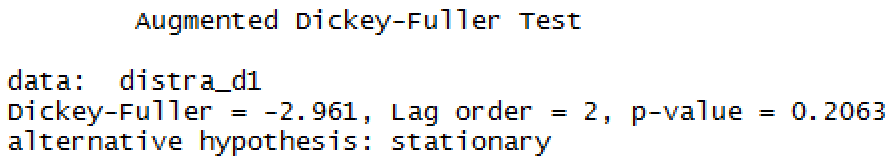


But has no improvement on residuals normality.

Distracted Driving.

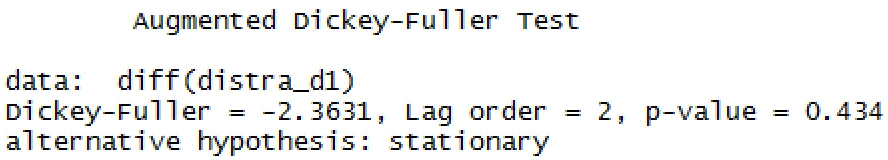
As mentioned earlier, distracted driving data is still not stationary. In addition, since other data are from 2011 first quarter until 2016 third quarter, while distracted driving data is from 2011 to 2015, we need to adjust the previous factors into the same period of time accordingly.

The number of crashes affected by distracted driving data, according to Augmented Dickey-Fuller Test, not only not stationary after first differencing,



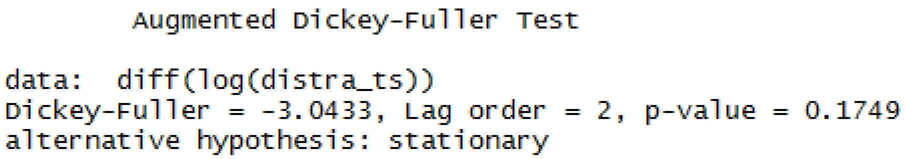
```
        Augmented Dickey-Fuller Test

data:  distra_d1
Dickey-Fuller = -2.961, Lag order = 2, p-value = 0.2063
alternative hypothesis: stationary
```

Also not stationary after second differencing,



```
        Augmented Dickey-Fuller Test

data:  diff(distra_d1)
Dickey-Fuller = -2.3631, Lag order = 2, p-value = 0.434
alternative hypothesis: stationary
```

It indicates the assumption that the data is moving average is questionable. So, let us look at log data differencing,

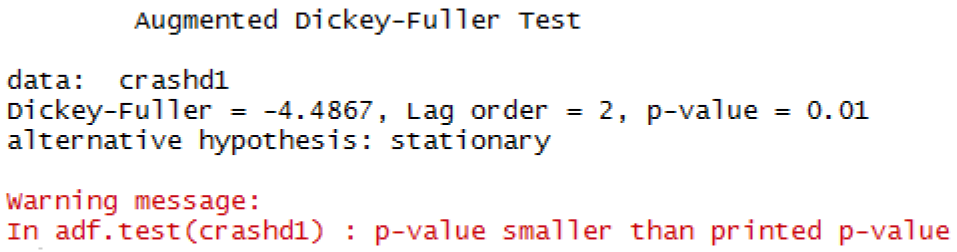


```
        Augmented Dickey-Fuller Test

data:  diff(log(distra_ts))
Dickey-Fuller = -3.0433, Lag order = 2, p-value = 0.1749
alternative hypothesis: stationary
```

Still not stationary with 0.05 significant level.

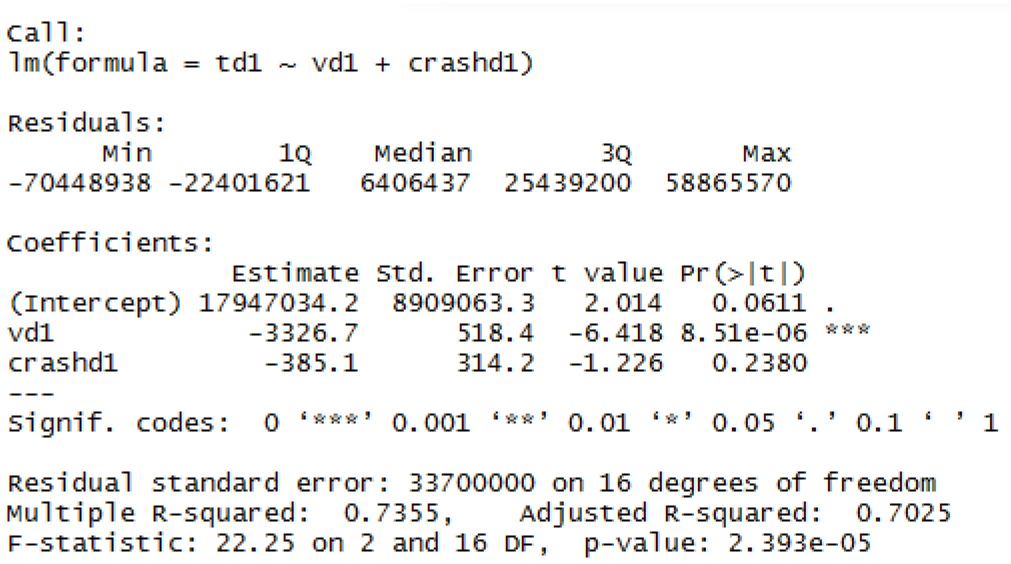
The second differencing after log still not stationary, and up to this point it is losing practical meaning.

Since the distracted driving data provided overall number of crashes, we decide to take this as a factor adding into the regression model. And fortunately the first differencing of the overall crashes data is stationary – since p-value smaller than 0.01.



```
        Augmented Dickey-Fuller Test

data:  crashd1
Dickey-Fuller = -4.4867, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(crashd1) : p-value smaller than printed p-value
```

Addition overall number of crashes increased adjusted R-squared from 0.67 to 0.7



```
Call:
lm(formula = td1 ~ vd1 + crashd1)

Residuals:
      Min        1Q    Median        3Q       Max 
-70448938 -22401621   6406437  25439200  58865570 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17947034.2  8909063.3   2.014   0.0611 .  
vd1            -3326.7      518.4  -6.418 8.51e-06 ***
crashd1         -385.1      314.2  -1.226   0.2380    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33700000 on 16 degrees of freedom
Multiple R-squared:  0.7355,	Adjusted R-squared:  0.7025 
F-statistic: 22.25 on 2 and 16 DF,  p-value: 2.393e-05
```

Although the number of crashes as factor itself is not significant.

Chapter 5: Conclusion and further research

Conclusions.

1. There is a relationship between car accidents loss amount and traffic volume, a correlation between the amount increased every quarter in traffic volume and the amount increased every quarter in loss cost.

$$Loss\ Cost_{\,t} - Loss\ Cost_{\,(t-1)} = -2658.7\,(Traffic\ volume_{\,t} - Traffic\ volume_{(t-1)}) + 14060931$$

In addition with overall number of crashes added,

$$Loss\ Cost_{year\ t} - Loss\ Cost_{year\ (t-1)} = -3326.7\,(Traffic\ volume_{year\ t} - Traffic\ volume_{year\ (t-1)}) - 385(Crashes_{year\ t} - Crashes_{year(t-1)}) + 17947034.2$$

2. The relationship maybe not linear, although the model doesn’t have signs to suggest a non-linear model so far, should try other non-linear models too base on the distribution of the data.
3. Collision plays very important role as representative of loss cost. It reflects the severity of the auto accidents very well. Maybe because it measures the damage of an accident by the dollar amount according to the same vehicle. Property damage amount varies and depending on what kind of vehicle the other party own in an accident.
4. Distracted driving may have something to do with losses amount, or it may be a different topic has other values we should look into in the future studies.

Future research.

1. To look into the trend of the loss amount, we should look into ARIMA model of the loss amount on its own.
2. Distracted driving has different kinds and may have effect on drivers and driving safety differently. Such as:

You can always see a traffic jam on the highway was because the other direction of the highway had an accident.

People wear headphones or playing loud music while driving a car.

People talk on the phone while driving the car. The cause of safety issues on this one actually is not because holding the phone. It is the talking part, so that even many newer cars all have “hands-free” device to allow people making phone calls while driving, I personally believe it doesn’t actually reduce the number of auto accidents caused by distraction.

Different kind of distraction make different amount of contributions to the auto accidents’ frequency and severity, to what degree each different kinds of distraction increase the likelihood of auto accidents is an interesting topic we may look into in the future.

3. Design and have an experiment on a specific road during a certain period of time, see the varying traffic volume and car accidents information. Because at different time during a day and different region of the country, the traffic volume vary all the time. This study only give us a big picture of how the traffic volume changing by month, maybe next time we can control the varying factors to see a simpler model how accidents vary with different traffic volume.
4. Take more factors into consideration.
5. There is a seasonal trend in traffic volume and loss amount. See if there is a way to remove the seasonal trend before using in the regression model or maybe consider non-linear seasonal model.

Reference

[1] Insurance Information Institute, October 2016. *Personal Automobile Insurance – More Accidents, Larger Claims Drive Costs Higher*.

[2] Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot. *Loss Models: From Data to Decisions 4th Edition.*

[3] Jonathan D. Cryer, Kung-Sik Chan. *Time Series Analysis With Applications in R*.

[4] Insurance Journal. *National Insurer Group Announces New Product, Fast Track Plus.*

http://www.insurancejournal.com/news/national/2006/03/13/66414.htm

[5] U.S Department of Transportation. National Highway Traffic Safety Administration. *Teen Distracted Driver Data 2015*

[6] Paul S.P. Cowpertwait, Andrew V. Metcalfe. *Introductory Time Series with R.*

[7] Lena Hiselius. *Estimating the Relationship Between Accident Frequency Homogeneous and Inhomogeneous Traffic Flow*

[8] Allstate Insurance Company. *What is a full coverage*.

https://www.allstate.com/tools-and-resources/car-insurance/what-is-full-coverage.aspx